

Consistency of the Maximum Likelihood Estimator of Evolutionary Tree

Arindam RoyChoudhury^{1*}

¹*Department of Biostatistics*

Columbia University, New York NY 10032, USA

* corresponding author

Maximum likelihood estimation (MLE) methods are widely used for evolutionary tree. As evolutionary tree is not a smooth parameter, the consistency of its MLE has been a topic of debate. It has been noted without proof that the classical proof of consistency by Wald holds for the MLE of evolutionary tree. Other proofs of consistency under various models were also proposed. Here we will discuss some shortcomings in some of these proofs and comment on the applicability of Wald's proof.

Key words: consistency, maximum likelihood, evolutionary tree,

phylogenetic tree, species tree

EXISTING PROOFS IN THE LITERATURE

The Proof by [11, 7, 5]

A proof was outlined by [11]. Then it was formalized by [7] and later restated in a simpler form by [5]. We will use the form provided by [5] to point out the subtle mathematical deficiency in the proof. The other versions of the proof suffer from the same deficiency.

The proof provided by [5] (pages 271-272) is as follows. Suppose that there are m possible character patterns x_1, x_2, \dots, x_m . The data consist of n independently observed patterns where x_i occurs n_i times. ($\sum_{i=1}^m n_i = n$.) Then the likelihood of tree (T) is

$$L(T) = \prod_{i=1}^m \Pr(x_i; T)^{n_i},$$

where $\Pr(x_i; T)$ is the probability of observing x_i under the tree-parameter value T . Thus,

$$\frac{1}{n} \log L(T) = \sum_{i=1}^m f_i^* \log p_i, \tag{1}$$

where $f_i^* = n_i/n$ and $p_i = \Pr(x_i; T)$.

Suppose that the true parameter value is T_0 and the probability of observing the i -th pattern is $p_i(0)$ under T_0 . Then for another set of

probability $p_i(j)$ (under another parameter value T_j):

$$\sum_{i=1}^m p_i(0) \log p_i(0) > \sum_{i=1}^m p_i(0) \log p_i(j)$$

(from Gibbs' inequality).

After this point the proof is supposed to proceed as follows,

$$\mathbb{E}\left(\frac{1}{n} \log(L(T_0))\right) - \mathbb{E}\left(\frac{1}{n} \log(L(T_j))\right) = \sum_{i=1}^m p_i(0) \log p_i(0) - \sum_{i=1}^m p_i(0) \log p_i(j).$$

Therefore,

$$\frac{1}{n} \log(L(T_0)) - \frac{1}{n} \log(L(T_j)) \rightarrow_{\text{a.s.}} \sum_{i=1}^m p_i(0) \log p_i(0) - \sum_{i=1}^m p_i(0) \log p_i(j) > 0$$

(using the Strong Law of Large Numbers.) Thus, there exist N_j , such that for all $n \geq N_j$,

$$L(T_0) > L(T_j) \tag{2}$$

with probability 1.

Now, if the parameter space was finite, then this would suffice. We could number the rest (other than T_0) of the parameters as T_1, \dots, T_p . Then for each $1 \leq j \leq p$, we could find N_j such that Eq. (2) holds. If we take

$$N \geq \max_{1 \leq j \leq p} N_j,$$

then for all $n \geq N$, $L(T_0)$ is the maximum likelihood with probability 1 (meaning that the MLE $\hat{T} = T_0$ with probability 1, and hence the MLE is consistent.) However, the parameter space is infinite, as there are infinitely many possible values for each branch length. To use a similar argument for infinite parameter space, we have to consider

$$N' \geq \sup_{x \in A} N_x, \quad (3)$$

where A is an indexing of the parameter space (excluding T_0). But, $\sup_{x \in A} N_x$ may be ∞ . Thus, our previous argument will not work.

[5] stated that some other conditions are needed to ensure the actual convergence of MLE to the true tree. He, however, did not provide those conditions. The intuitive argument of [11] does not address the mathematical problem described above.

The Proof by [8]

[8] numbered the tree-topologies as $1, 2, \dots, K$. The proof by [8] treats the numbered topologies and other numeric parameters as points in Cartesian space, and attempts to establish that the parameter space thus constructed is closed. It argues that as each parameter θ' (including the numeric tree-topology parameter) can be transformed so that $0 \leq \theta' \leq c$ for

a constant c , the transformed parameter space is closed. However, this argument is not enough to prove that a set is closed. [8] considered only completely bifurcating trees in the parameter space. Thus for a given internal branch the set of all possible branch lengths does not include 0. Therefore, the parameter space is not closed.

Including all trees in the parameter space may appear to be an easy remedy to this. But, that leads to some inherent difficulties, as described below.

Any multifurcated tree-topology Y can be obtained by collapsing one or more internal branches of at least two distinct completely bifurcating tree-topologies Y_1 and Y_2 . Therefore, one can have two sequence of trees S_1 and S_2 , having tree-topologies Y_1 and Y_2 respectively, such that they converge to a common limiting tree with tree-topology Y . Therefore, to be closed, Y has to have same numeric value as both Y_1 and Y_2 , which is impossible under the numbering scheme of [8].

An identifiability problem with the proof was also discussed by [1]. (Identifiability is a requirement for consistency as unidentifiable parameter space could lead to serious estimation problems; see, for example, [6].)

Applicability of the proof by [10]

As mentioned before, [4] noted that the proof by [10] can be used to proof the consistency. [9] debunked the claim of some works (such as [12] and [3]) that argued that the differentiability and continuity of likelihood (as a function of the parameters) is a requirement in the proof by [10].

However, Assumption 7 (the parameter space has to be a closed subset of a Cartesian space) of [10] appears not to hold. (For a detailed analysis of the assumptions of Wald, 1949 in the context of genetic data, see Rogers, 2001).

[10] noted that Assumption 7 is unnecessarily binding and stated an alternative Assumption 9 in [10]. The alternative involves a condition that a metric has to be defined in the parameter space, under which all closed and bounded sets are compact (Assumption 9(i,iv)). (The other conditions in Assumption 9 will be satisfied by any reasonable model.)

There are two kinds of parameters associated with a tree: branch-specific (such as branch length) and non-branch-specific (the overall parameters that are not associated with a particular branch). The set of all possible values of the branch-specific parameters can be shown to have one-to-one correspondence with a complete real metric space, where the natural metric conforms to the natural concept of distance in evolutionary trees ([2]). (In a real valued complete metric space all closed and bounded

sets are compact; moreover, the definition of closed set would be preserved from the tree-space to the metric space as the concept of distance is preserved.) Therefore, if all the non-branch-specific parameters also form a real complete metric space, then the MLE is consistent for evolutionary tree.

DISCUSSION

We have pointed out mathematical shortcomings in some of the existing proofs of consistency of MLE of evolutionary tree. We have also established that Wald's proof of consistency [10] can indeed be used to prove the consistency of MLE of evolutionary tree as suggested by [4].

A simulation-based verification of consistency is not feasible for trees with a large number of taxa. As trees with a larger number of taxa can potentially introduce structures that are not present in fewer-taxa trees, a simulation-based verification of consistency in the latter is not enough to assert the consistency in the former. Therefore, a theoretical proof of consistency (such as the one noted in this work) is required to ensure that the MLE of evolutionary tree with any number of taxa is indeed consistent.

ACKNOWLEDGMENTS

The author is grateful to Joseph Felsenstein, John Wakeley and Adam Seipel for their valuable suggestions.

REFERENCES

- [1] E. S. Allman, C. Ané, and J. A. Rhodes. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Adv. Appl. Prob.*, 40:228–249, 2008.
- [2] L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Adv. Appl. Math.*, 27:733–767, 2001.
- [3] J. S. Farris. Likelihood and inconsistency. *Cladistics*, 15:199–204, 1999.
- [4] J. Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.*, 25:471–492, 1973.
- [5] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004.
- [6] F. A. Matsen and M. Steel. Phylogenetic mixture on a single tree can mimic a tree of another topology. *Syst. Biol.*, 56:767–775, 2007.

- [7] J. S. Rogers. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Syst. Biol.*, 46:354–357, 1997.
- [8] J. S. Rogers. Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Syst. Biol.*, 50:713–722, 2001.
- [9] D. L. Swofford, P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.*, 50:525–539, 2001.
- [10] A. Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.*, 20:595–601, 1949.
- [11] Z. Yang. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst. Biol.*, 43:329–342, 1994.
- [12] Z. Yang. Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.*, 42:294–307, 1996.